Chapitre 7

Annotations et mesures

À partir des enregistrements de parole spontanée recueillis, nous souhaitons observer où les locuteurs ont tendance à placer leurs pauses et comment ils produisent l'accent lexical. En outre, s'agissant de parole conversationnelle, un certain nombre de traitements sont nécessaires en amont pour pouvoir effectuer ces mesures. Nous tenons également à ce que l'ensemble de ces traitements soit fait de manière automatique, afin de voir si ce type de mesures peut être effectué sans intervention manuelle.

Ce chapitre présente les différents traitements réalisés et les métriques utilisées pour les évaluer, ainsi que les mesures effectuées pour analyser les patterns de pauses et d'accentuation lexicale.

7.1 Identification du locuteur

Tout d'abord, les enregistrements de parole que nous voulons analyser sont des conversations spontanées entre plusieurs locuteurs, il est donc avant tout nécessaire d'identifier qui parle quand, de manière à pouvoir relier chaque phénomène de parole au bon locuteur.

À l'issue de ce module de traitement, nous souhaitons obtenir des segments de parole mono-locuteurs qui correspondent plus ou moins aux tours de parole. La principale contrainte qui se pose ici est de savoir jusqu'où nous tolérons les réactions de l'interlocuteur lorsqu'elles ne coupent pas la parole du locuteur. En effet, si l'outil de segmentation est trop sensible, la moindre réaction de l'interlocuteur risque de mettre fin au segment de parole en cours, résultant en de nombreux segments courts et non-terminés, ou commençant au milieu d'un énoncé. À l'inverse, si l'outil n'est pas assez sensible, les segments risquent de contenir beaucoup de parole de l'interlocuteur, qui

pourraient alors être analysée par erreur comme provenant du locuteur. En outre, la solution de supprimer a posteriori les passages de l'interlocuteur risquent d'affecter également la parole du locuteur, notamment lorsqu'il y a chevauchement de parole. Il faudrait pouvoir isoler la parole de chaque locuteur afin de pouvoir en conserver qu'une, même dans les cas de chevauchements.

Il est important de vérifier la qualité de la segmentation de parole et l'annotation en locuteur pour s'assurer que les mesures effectuées par la suite sont attribuées au bon locuteur. La précision de la reconnaissance étant dépendante du type d'enregistrements utilisés (qualité audio, nombre et voix des locuteurs, organisation des tours de parole etc.), nous proposons de tester le module sur un échantillon du corpus CLES manuellement annoté en locuteurs. Cet échantillon consiste en 20 enregistrements (3 h, 40 locuteurs, 2 locuteurs par enregistrements) segmentés automatiquement avec Whisper puis annotés manuellement en locuteurs.

La comparaison de l'annotation automatique et de l'annotation manuelle sera effectuée indépendamment pour chaque locuteur, en observant ⓐ la proportion de segment correspondant au locuteur cible, ⓑ la proportion de segment qui correspond au mauvais locuteur, et ⓒ la proportion de segment du locuteur cible manquée. À partir de ces valeurs, nous pouvons calculer un score de précision ($\frac{ⓐ}{ⓑ}$) et de rappel ($\frac{ⓐ}{ⓐ+ⓒ}$), et ainsi quantifier la proportion d'erreur par locuteur (qui impactera inéluctablement les résultats individuels), et la proportion de parole correctement annotée.

ILLUSTRATION ICI

7.2 Reconnaissance et alignement de la parole

L'étape suivante consiste à transcrire la parole des locuteurs et de d'aligner la transcription au signal, au moins au niveau du mot, de manière à pouvoir localiser les pauses en contexte, ainsi que les mots sur lesquels nous effectuerons les mesures acoustiques d'accentuation.

La première question qui se pose ici est de savoir s'il vaut mieux transcrire l'ensemble de la conversation puis de segmenter en locuteurs, ou bien de segmenter d'abord, puis de transcrire segment par segment. Dans le premier cas, le système de reconnaissance dispose de l'ensemble de la conversation et donc du contexte global, pouvant améliorer la précision de la reconnaissance ; en contrepartie, la conversation est longue (environ 10 min) et plusieurs locuteurs se partagent la parole avec d'éventuels chevauchements. Il semble de plus en plus envisageable de choisir la première option car les systèmes de reconnaissance de la parole gèrent de mieux en mieux la reconnaissance des dialogues spontanés ; toutefois, au moment de faire ce choix, en

2021, nous optons pour la deuxième option, qui nous semble plus raisonnable : segmenter d'abord puis transcrire chaque segment indépendamment.

Pour évaluer la qualité de la reconnaissance de la parole, nous avons calculé le taux d'erreur mot *(word error rate, WER)* sur des corpus de parole plus ou moins contrôlée. D'abord sur des phrases porteuses élicitées par des locuteurs anglophones natifs, japonophones et coréanophones (5 h22 min, 4 799 phrases, 54 locuteurs) ; puis sur des textes lus par des locuteurs natifs et japonophones (34 h, 954 textes, 57 locuteurs) ainsi que des locuteurs francophones (3 h45 min, 1 texte commun, 148 locuteurs) ; et enfin sur de la parole spontanée transcrite manuellement issue du corpus CLES mentionné dans la section précédente (3 h, 40 locuteurs). Le WER est calculé en faisant le somme des substitutions, délétions et insertions, le tout divisé par le nombre total de mots dans le texte de référence, donnant un pourcentage d'erreur de reconnaissance. En complément du WER, nous analyserons également le nombre de substitutions, délétions et insertions.

Dans le cas des phrases porteuses et des textes lus, nous avons pris pour référence le texte source, en partant du principe que les lecteurs ont lu ce qui leur était demandé de lire. Pour la parole spontanée, les transcriptions ont été obtenues automatiquement avec Whisper puis corrigées manuellement.

Pour évaluer la qualité de l'alignement au niveau du mot, nous proposons d'adapter la méthode employée dans la section précédente pour comparer l'alignement cible à un alignement de référence. Nous disposons de deux enregistrements alignés au niveau du mot, qui pourront servir de référence. Ces enregistrements proviennent de l'étude de Frost et al. (2024), il s'agit de l'enregistrement de deux enseignants francophones faisant leur cours en anglais (3 min48 s et 3 min34 s). Les enregistrements ont d'abord été alignés automatiquement avec WebMAUS (Kisler et al., 2017), puis corrigés manuellement par un phonéticien. Nous utiliserons cet alignement comme référence. Pour chaque enregistrement, nous calculerons la proportion d'alignement correspondant entre l'hypothèse et la référence, ainsi que la proportion d'erreur d'alignement spécifique aux mots cibles sur lesquels porteront les analyses acoustiques.

7.3 Détection des noyaux syllabiques

Pour effectuer les mesures d'accentuation lexicale, il est encore nécessaire de segmenter les mots en syllabes et identifier les noyaux syllabiques sur lesquels porteront les analyses acoustiques. Nous envisageons deux méthodes pour localiser les syllabes : une méthode acoustique et une méthode phonologique. La méthode acoustique consiste à se baser sur les pics d'intensité. Chaque syllabe d'un mot consiste en principe en un pic d'intensité, nous pourrions donc considérer chaque pic à l'intérieur

d'un mot comme un noyau de syllabe. La méthode phonologique consiste quant à elle à utiliser un alignement forcé de chaque phonème correspondant au mot, à partir d'un dictionnaire phonologique, puis de considérer chaque intervalle vocalique comme noyau syllabique. Dans le premier cas de figure, les noyaux sont donc représentés par des points (maximums d'intensité), et dans le second, ce sont des intervalles de durée vocalique.

Si la méthode phonologique permet d'obtenir autant d'intervalles vocaliques que de syllabes attendues pour un mot donné (puisque l'alignement se base sur un dictionnaire phonologique), la méthode acoustique permet de d'estimer le nombre et la position des syllabes quelque soit le mot et la façon dont il est prononcé. Toutefois, nous n'avons pas encore formalisé de méthode pour vérifier si les noyaux acoustiques correspondent effectivement à un noyau vocalique, ou si les noyaux non-détectés le sont effectivement à cause d'une élision de syllabe par le locuteur. Nous nous contenterons dans un premier temps de compter le nombre de mots dont le nombre de syllabes acoustiques correspond au nombre de syllabe attendu dans un dictionnaire phonologique de référence.

7.4 Annotation des pauses

Comme indiqué par plusieurs études précédentes (de Jong, 2016 ; Kahng, 2018 ; Kallio et al., 2022 ; Suzuki & Kormos, 2020, entre autres), la distribution des pauses est dépendante de la syntaxe de l'énoncé. On aura ainsi tendance à observer les pauses en frontière de constituants plutôt qu'à l'intérieur de ceux-ci ; et plus le nombre de pauses intra-constituant est élevé, plus la parole a tendance à être jugée disfluente. Nous souhaitons donc localiser les pauses et les catégoriser en fonction de leur contexte syntaxique. Comme Fauth et Trouvain (2018), nous entendrons par « pause » toute interruption de parole, qu'il s'agisse de pause pleine ou silencieuse, faux départ, répétitions ou allongements.

À ce stade, nous disposons de l'alignement des mots au signal de parole. Nous pouvons donc par extension localiser dans la chaîne de texte les silences ou ce qui peut constituer une pause pleine. Chaque intervalle séparant deux mots sera considéré comme pause potentielle. Il sera ensuite possible de définir un seuil de durée minimum pour considérer ou non ces intervalles comme des pauses. À l'avenir, nous souhaitons toutefois moduler ce seuil en fonction du débit de parole (cf. chapitre 14).

Pour étiqueter les pauses en fonction de leur position syntaxique, nous proposons d'effectuer une analyse syntaxique par constituants, pour délimiter et hiérarchiser l'énoncé en propositions et en syntagmes. Chaque intervalle sera ainsi annoté de son contexte gauche et droit : la catégorie du mot qui précède et qui suit, le consti-

tuant le plus grand qui se termine, le nombre de mots qu'il contient et sa profondeur syntaxique (estimée à partir du nombre de constituants en cours), le constituant le plus grand qui commence, son nombre de mots et sa profondeur syntaxique. L'étiquette du constituant pourra ensuite être interprétée en frontière de proposition, de syntagme ou de mot à partir des catégories de constituants de PennTree Bank, donnée en Annexe A.

7.4.1 Analyses

Dans cette étude, nous considérons un seuil de durée minimum fixe de 180 ms pour prendre en compte les pauses brèves tout en évitant les phénomènes de coarticulation (Heldner & Edlund, 2010). Pour permettre une meilleure comparabilité de nos résultats avec les études précédentes, nous considérerons également le seuil de 250 ms, plus commun dans domaine de l'évaluation de la fluence en L2. Un seuil de durée maximum de 2 s est également paramétré de manière à ignorer les pauses très longues, pouvant résulter d'erreurs d'alignement.

Grâce à l'étiquetage présenté ci-dessus, nous pouvons catégoriser chaque frontière de mot en fonction du type de frontière syntaxique : inter-proposition (si une proposition se termine ou commence), inter-syntagme (si un syntagme se termine ou commence), ou à défaut, intra-syntagme (si aucune frontière de constituant n'est présente). Par extension, nous pouvons donc catégoriser les pauses en fonction du type de frontière sur laquelle elles interviennent. On pourra alors comparer les groupes de locuteurs (B1, B2, francophones, japonophones ou anglophones natifs) en fonction de la fréquence des pauses produites et de leur distribution syntaxique.

On trouve différentes mesures de fréquence des pauses dans la littérature : le nombre de pauses par minute, par mot, par syllabe, ou encore par tour de parole. La fréquence des pauses par minute est influencée par le débit de parole : plus le locuteur parle vite, plus le nombre de mots par minute augmente et par conséquent le nombre de pauses potentielles, bien que cela puisse paraître contre-intuitif. Pour neutraliser l'influence du débit de parole, nous choisirons de calculer la fréquence des pauses par mot, ou plus exactement par token issu de la phase de transcription et d'alignement. Pour la fréquence des pauses en fonction de leur position, nous calculerons d'abord la fréquence des pauses $F_{p,i}$ par type de frontière syntaxique i (inter-proposition (BC), inter-syntagme (BP) et intra-syntagme (WP)) :

$$F_{p,i\in\{BC,BP,WP\}} = \frac{N_{p,i}}{N_i} \tag{7.1}$$

avec $N_{p,i}$ le nombre de pauses p par catégorie i, et N_i le nombre de frontières

syntaxique de type i. La valeur obtenue indique par exemple à quelle fréquence deux propositions sont séparées par une pause chez un locuteur donné. Nous compléterons cette mesure par la proportion de pauses P_i de chaque catégorie i :

$$P_{i \in \{BC,BP,WP\}} = \frac{N_{p,i}}{N_p} \tag{7.2}$$

avec N_p le nombre total de pauses, toutes catégories confondues. Cette valeur indique par exemple la proportion de pauses intra-syntagmes chez un locuteur.

Comparer les groupes de locuteurs revient à comparer les scores obtenus par locuteur. Étant donné le nombre parfois limité de locuteurs (notamment pour le corpus japonophone et anglophone) et la non-normalité des distributions, la comparaison se fera au moyen du test de rangs non-paramétrique Wilcoxon-Mann-Whitney (Bauer, 1972). Ce test se concentre sur la différence de tendance générale entre deux distributions, mais a pour avantage d'être robuste à la taille et au type de distribution des données. Nous nous baserons sur les résultats de ce test pour vérifier la significativité de la différence entre les distributions. Nous indiquerons également la tendance centrale de chaque distribution en indiquant leur valeur médiane, ainsi que la taille d'effet pour quantifier le degré de différence entre elles. Pour cela, nous proposons de calculer le delta de Cliff (Cliff, 1993), qui indique à quel point les valeurs d'une distribution A sont supérieures ou inférieures à celles de la distribution B. Le delta obtenu varie entre -1 et 1, 0 indiquant que les deux distributions sont identiques, 1 indiquant que toutes les valeurs de la première distribution sont supérieures à celles de la deuxième. Nous utiliserons les seuils d'interprétation de Romano et al. (2006) : la différence est grande à partir de $0{,}474$, moyenne à partir de $0{,}33$, et petite à partir de $0{,}147$; inférieure à cette valeur, la différence est négligeable.

7.4.2 Score de distribution syntaxique

Nous proposons de calculer un score de distribution syntaxique des pauses (SDS) qui représente en une seule valeur le niveau syntaxique auquel les pauses ont tendance à survenir chez un locuteur, et ce indépendamment du nombre de pauses produites. Le calcul est effectué comme suit : pour un échantillon de parole donné, on compte le nombre de pauses de chaque catégorie (inter-proposition (BC), inter-syntagme (BP) et intra-syntagme (WP)) que l'on pondère de manière à favoriser les pauses de haut niveau et pénaliser celles de bas niveau. Enfin, on normalise par le nombre total de pauses N. Ce calcul peut se noter de la façon suivante :

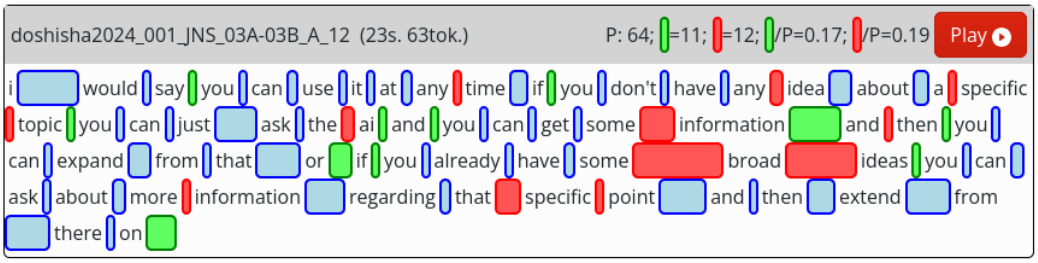

FIG. 7.1 : Transcription d'un segment audio indiquant chaque type de frontière syntaxique par une couleur : vert pour inter-proprosition, bleu pour inter-syntagme et rouge pour intra-syntagme. La longueur des intervalles est proportionnelle à leur durée ; seule les plus longs sont considérés comme des pauses par PLSPP (cliquer ici pour accéder à la visualisation en ligne)

$$SDS = \sum_{i \in BC,BP,WP} (p_i \cdot w_i) = \frac{N_{BC} \cdot w_{BC} + N_{BP} \cdot w_{BP} + N_{WP} \cdot w_{WP}}{N_p} \quad (7.3)$$

Nous proposons de fixer w_{BC} à 1, w_{BP} à 0,5 et w_{WP} à -1, de manière à faire varier le score entre -1 et 1. La présence de pauses inter-proposition et inter-syntagme participeront à élever le score, avec plus de poids pour les premières, tandis que les pauses intra-syntagme tireront le score vers le bas. Plus le score est haut, plus les pauses ont tendance à être placées en frontières de haut niveau. Un score négatif indique que la majorité des pauses est placée intra-syntagme, ce qui semble toutefois peu probable.

7.4.3 Amélioration de l'approche

Selon nous, considérer les pauses en fonction de leur position vis-à-vis des propositions ou des syntagmes présente deux limitations importantes. La première est le fait que le nombre de frontières intra-syntagmes est assez limité en anglais, et réduit ainsi la probabilité d'y trouver une pause. La figure 7.1 illustre cet état de fait : elle présente un segment du corpus CLES-JP où chaque frontière syntaxique est colorée en fonction de son niveau. On y voit seulement 12 frontières intra-syntagme (en rouge), contre 41 inter-syntagmes (bleues). La deuxième limitation est le fait que toutes les frontières qui ne sont ni inter-proposition ni intra-syntagme sont considérées au même niveau "inter-syntagme", alors qu'il y a en réalité toute une hiérarchie de syntagmes imbriqués les uns dans les autres – ce qui est, par ailleurs, également le cas pour les propositions. Nous proposons donc une nouvelle approche pour contourner ces limitations.

Au lieu de considérer des niveaux de frontières syntaxiques en fonction de leur

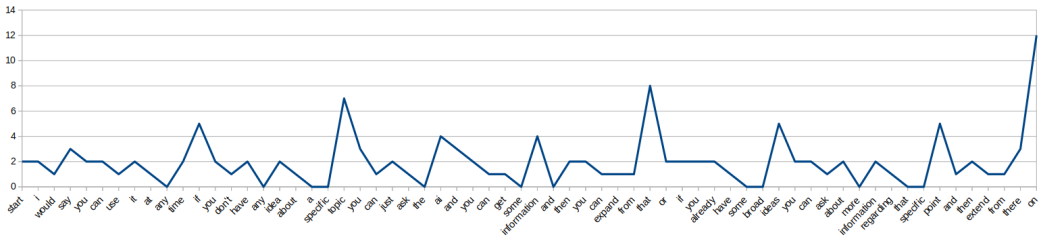

FIG. 7.2 : Nombre de constituants se fermant ou s'ouvrant après chaque mot

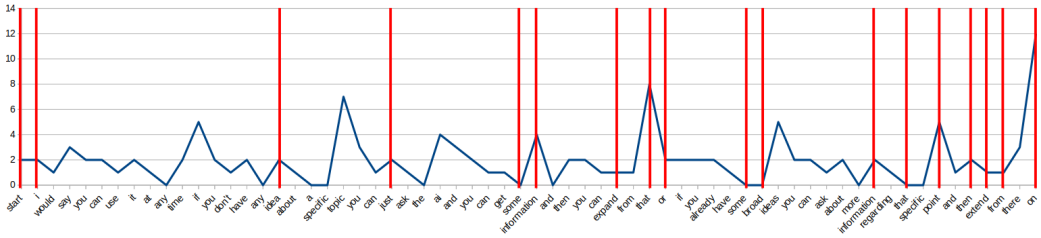

FIG. 7.3 : Même figure avec position des pauses (en rouge, seuil de durée minimale à 180 ms)

type, il s'agit de les considérer relativement aux autres. Plus le nombre de constituants qui s'ouvrent ou se ferment est élevé à un endroit donné, plus la frontière est importante. Quelque soit la nature de la frontière, c'est leur concomitance qui définit leur importance. Ainsi, la fermeture de trois syntagmes imbriqués les uns aux autres donnera une frontière plus importante que la fermeture d'un seul syntagme. On peut alors calculer une valeur représentant ce degré de frontière – par exemple la somme des constituants qui se ferment ou qui s'ouvrent après chaque mot. La figure 7.2 représente cette valeur pour le même segment que la figure précédente. On y distingue des pics qui semblent correspondre à des positions naturelles de pauses dans l'énoncé. Plus la valeur est basse en revanche, moins une pause à cet endroit semble probable.

On peut calculer le score de distribution syntaxique en remplaçant les catégories de pauses par des seuils d'importance de frontière. Nous proposons de définir arbitrairement trois seuils de la manière suivante : *high* pour une frontière d'importance 4 ou plus, *medium* pour 2 ou 3, *low* pour 0 ou 1. Le calcul est le même que ci-dessus : on fait une somme pondérée de la fréquence de pauses par niveau, avec les mêmes poids w_{high}, w_{medium} et w_{low} de 1, 0,5 et -1.

En résumé, nous proposons d'effectuer les mesures suivantes :

- **Analyse globale :** nombre de pauses par token et durée moyenne des pauses ;
- **Analyse structurelle :** fréquence des pauses par type de frontière syntaxique ($F_{p,i}$), proportion des pauses par catégorie (P_i), score de distribution syntaxique basé sur les propositions et les syntagmes, score de distribution syntaxique basé

sur le niveau d'importance des frontières.

7.4.4 Évaluation de l'étiquetage

Pour évaluer la précision de la détection et de l'annotation des pauses, nous proposons de les comparer aux annotations manuelles du corpus Mareková, développé à l'université Constantine le Philosophe à Nitra (Slovaquie), et l'Institut d'Informatique de Bratislava (REFERENCE). Le corpus est composé de 72 dialogues de 24 binômes d'étudiants de langue maternelle slovaque, lors d'un jeu de rôle où les locuteurs sont amenés à décrire un trajet sur une carte. Le corpus totalise 8 h18 min de parole spontanée conversationnelle, et est entièrement transcrit est annoté manuellement en pauses inter- et intra-propositionnelles. Nous proposons de comparer le nombre et la catégorie des pauses annotées avec les résultats de nos annotations automatiques. Plus concrètement, il s'agira de calculer le score de précision et de rappel de l'annotation automatique à partir des annotations de référence, et tenter de quantifier et expliquer les erreurs de détection et d'étiquetage.

Nous n'avons malheureusement pas trouvé de corpus annoté manuellement en syntagmes.

7.5 Annotation de l'accent lexical

L'accent lexical joue un rôle important pour la segmentation du flux de parole et l'accès lexical (Cutler, 2015 ; Cutler & Jesse, 2021). La qualité de sa réalisation est corrélée avec les jugements de compréhensibilité des auditeurs, et ce pour les débutants comme pour les locuteurs de niveau avancé (Isaacs & Trofimovich, 2012 ; Saito et al., 2015). Nous ne disposons pas à ce jour de système d'évaluation de la précision de l'accent lexical adapté à la parole spontanée (Saito et al., 2022). Nous tenterons d'apporter une solution de traitement possible.

L'accentuation lexicale en anglais est réalisée par une combinaison de facteurs prosodiques et segmentaux qui font varier la qualité de la syllabe. La syllabe accentuée est en général plus haute en F_0, en intensité, et de durée plus longue que les syllabes non-accentuées, qui ont tendance au contraire à être réduites (plus basse en F_0, en intensité, et plus courte en durée). Au niveau segmental, la voyelle accentuée est pleine et parfois diphtonguée, tandis que la voyelle réduite est centralisée et tend vers le phonème /ə/. Par ailleurs, les mots lexicaux (noms, adjectifs, verbes, adverbes) ont tendance à être accentués, alors que les mots grammaticaux ont tendance à être réduits.

Nous proposons dans un premier temps de nous concentrer sur l'accentuation

des mots polysyllabiques lexicaux. À ce stade des traitements, nous disposons d'un alignement des mots et de leurs syllabes au signal de parole, ainsi que la catégorie grammaticale issue de l'analyse morphosyntaxique. Nous sommes donc en principe en mesure d'identifier le patron accentuel attendu pour chaque mot du corpus, en recourant à un dictionnaire phonologique de référence. Pour chaque mot polysyllabique lexical, nous proposons de mesurer le degrés de proéminence syllabique à partir de la F_0, de l'intensité et de la durée de chaque syllabe. La syllabe qui obtient le score maximum sera considérée comme la syllabe accentuée, et les autres seront pour l'instant considérées comme non accentuées (modèle binaire). Chaque dimension prosodique sera par ailleurs normalisée par locuteur et représentée en centile. Ainsi, une F_0 de 50 indiquera une valeur médiane pour le locuteur en question, comparable à la valeur 50 de n'importe quel autre locuteur. Plus la valeur tend vers 100, plus la F_0 est élevée. Cette méthode de normalisation permet de tenir compte de la distribution des mesures pour chaque locuteur, tout en permettant de comparer les valeurs entre elles (50 représente la valeur médiane pour tous les locuteurs sur toutes les dimensions). En contrepartie, il est nécessaire d'avoir suffisamment de mesures pour chaque locuteur, sans quoi des centiles différents peuvent renvoyer aux mêmes valeurs absolues.

Deux scores seront ensuite calculés par locuteur : un score de position de l'accent, représentant le pourcentage de mots pour lesquels la syllabe proéminente correspond à la syllabe qui porte l'accent lexical selon le dictionnaire de référence (on pourrait donc dire que l'accent est correctement positionné) ; et un contraste prosodique moyen calculé à partir de la différence entre la valeur prosodique de la syllabe censée être accentuée et la moyenne des autres syllabes, sur l'ensemble des mots produits par un locuteur. Ce contraste pourra être calculé globalement (moyenne des trois dimensions prosodiques) ou par dimension. Il indique ainsi à quel point la syllabe accentuée se démarque acoustiquement des autres. La valeur obtenue varie entre -100 et +100, 0 indiquant qu'il n'y a pas de différence prosodique entre la syllabe accentuée et les autres syllabes, +100 indique un contraste maximum positif, -100 indique un contraste maximum négatif, signalant que la proéminence se situe sur une autre syllabe que celle censée être accentuée.

Nous effectuerons les mesures suivantes pour chaque groupe de locuteurs :

- Nombre de mots, nombre de mots polysyllabiques lexicaux, nombre de mots annotés ;
- Proportion de mots par catégorie grammaticale et par nombre de syllabes ;
- Proportion de mots selon la position de l'accent lexical (attendu) et selon la position de la syllabe proéminente produite (réalisé) ;
- Score de position de l'accent par locuteur ;

- Contraste prosodique moyen par locuteur, moyen et par dimension (F_0, intensité et durée) ;

Comment savoir si la syllabe proéminente identifiée par le système correspond effectivement à la syllabe accentuée perçue par l'auditeur ? Nous proposons quatre approches différentes pour aborder cette question :

a) Demander à des locuteurs anglophones natifs de noter manuellement les syllabes qu'ils perçoivent accentuées dans des enregistrements de locuteurs non-natifs et comparer avec les annotations automatiques ;

b) Demander à des locuteurs natifs et non-natifs où doit être placé l'accent primaire sur une série de mots cibles, puis comparer leur conscience accentuelle avec leur production ;

c) Annoter automatiquement des enregistrements de parole plus ou moins contrôlée produite par des locuteurs natifs ;

d) Comparer les résultats obtenus sur différents types de parole avec 1) la méthode acoustique et 2) la méthode phonologique d'identification des noyaux syllabiques, pour voir si les mesures sont cohérentes et si les mêmes tendances générales sont observées.

a) Évaluation perceptive par des locuteurs natifs

La première approche consiste à vérifier si les annotations automatiques d'accentuation lexicale sont cohérentes avec le jugement d'auditeurs natifs. En d'autres mots : est-ce que les anglophones natifs entendent l'accent au même endroit que la machine ? Nous avons recruté 10 évaluateurs anglophones natifs à qui nous avons fait annoter manuellement 6 enregistrements de locuteurs japonophones. Les évaluateurs sont originaires des États-Unis et vivent dans la région de Tōkyō depuis plus de 5 ans au moment de l'expérimentation, ils sont donc habitués à l'influence du japonais sur la prononciation de l'anglais, mais aucun d'entre eux n'est enseignant d'anglais. En ce qui concerne les enregistrements, six élèves entre 9 et 11 ans d'une école primaire privée de la préfecture de Kyōto ont été enregistrés pendant une récitation de texte. Le texte est une description d'un personnage historique de 300 mots, commun à l'ensemble des locuteurs, et qui a fait l'objet d'un entraînement préalable. La transcription du texte avec les mots à évaluer mis en relief est fournie aux évaluateurs au format papier, en 6 exemplaires, et les évaluateurs doivent noter à la main la position de l'accent qu'ils perçoivent pour chaque enregistrement écouté sans contrainte d'écoute, dans un ordre aléatoire.

Pour un mot donné, si l'accent est noté sur une syllabe qui ne correspond pas à la syllabe censée porter l'accent primaire d'après le dictionnaire de référence, on compte une erreur. Après avoir calculé le taux de corrélation inter-annotateur, nous avons comparé le nombre d'erreurs relevées par évaluateur et par locuteur, et l'avons comparé au nombre d'erreurs identifiées automatiquement. Par ailleurs, un score d'accentuation S_w a été calculé pour chaque mot à partir des valeurs prosodiques mesurées par le système, puis comparé à un second score indiquant la moyenne des jugements humains pour le mot en question. L'équation 7.4 détaille le calcul effectué pour obtenir le score d'accentuation :

$$S_w = \frac{P_{s,w}}{P_{s,w} + \overline{P_{u,w}}} \tag{7.4}$$

où w est le mot courant, $P_{s,w}$ correspond à la valeur prosodique de la syllabe accentuée attendue (moyenne des centiles de F_0, d'intensité et de durée), et $\overline{P_{u,w}}$ la valeur prosodique moyenne des autres syllabes du mot. On obtient alors une valeur en 0 et 1, 0.5 indiquant aucun contraste entre la syllabe accentuée et les autres syllabes, et 1 indiquant un contraste positif maximal. Les résultats obtenus sont détaillés en section 3.1.

b) Annotation automatique et conscience phonologique

La deuxième approche a consisté à comparer l'annotation automatique et le jugement de position théorique de l'accent par les mêmes locuteurs. Une liste de 57 mots cibles a été enregistrée dans des phrases porteuses par 12 locuteurs anglophones natifs, 14 locuteurs japonophones et 11 locuteurs coréanophones, puis annotée automatiquement avec notre système. En parallèle, ces mêmes locuteurs ont passé un test de conscience phonologique, lors duquel il leur était demandé d'indiquer la voyelle qui, selon eux, porte l'accent primaire, sur les mêmes mots cibles. Les mots sélectionnés consistent en 19 triplets composés d'un verbe à 3 syllabes portant l'accent sur l'initiale (ex. *dominate*), sa forme en *-ing* (accent primaire sur l'initiale, ex. *dominating*), et son dérivé substantif en *-ion* (accent primaire sur la 3ème syllabe, ex. *domination*). Cette approche permet de vérifier si la syllabe proéminente identifiée automatiquement correspond à la syllabe considérée accentuée par les locuteurs, indépendamment d'une référence prescriptive externe.

Un taux de correspondance entre l'accent théorique et l'annotation automatique a été calculé pour chaque groupe de locuteurs et chaque item du triplet. Une observation des mesures acoustiques de chaque syllabe a également permis d'étudier le poids donné à l'accent secondaire vis-à-vis de l'accent primaire. Les résultats obtenus sont détaillés en section 3.2.

c) Annotation de parole produite par des locuteurs natifs

La troisième approche a consisté à considérer les locuteurs anglophones natifs comme une référence en termes d'accentuation lexicale, en partant du principe qu'ils accentueront systématiquement la syllabe censée porter l'accent primaire. Nous avons commencé avec de la parole lue en studio par des professionnels de la voix, enregistrés dans le cadre de la constitution d'un manuel scolaire d'anglais, et donc avec pour objectif de représenter un modèle d'anglais idéal. Par la suite, nous avons annoté des enregistrements de parole spontanée conversationnelle issus du corpus CLES.

Dans les deux cas, nous avons calculé la proportion de mots accentuée conformément au dictionnaire de référence, et le degré de contraste prosodique mesuré sur chaque dimension entre la syllabe accentuée attendue et les autres syllabes du mot. Les résultats obtenus en parole lue en studio sont présentés en section 3.3 ; ceux obtenus en parole spontanée sont présentés dans le chapitre 9.

d) Comparaison méthode acoustique *vs.* méthode phonologique

Enfin, nous avons souhaité observer comment varient les annotations automatiques selon que les noyaux syllabiques sont extraits de manière acoustique ou phonologique. Pour ce faire, nous avons comparé les résultats obtenus sur plusieurs corpus (parole lue, locuteurs natifs et non-natifs), en termes de proportion de mots correctement accentués par niveau et par locuteur, ainsi que de contraste prosodique observé sur chacune des 3 dimensions. Les résultats sont présentés section 3.4.